ELSEVIER

Cognitive Science 27 (2003) 453–489

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Inferring causal networks from observations and interventions

Mark Steyvers [a,*], Joshua B. Tenenbaum [b,1], Eric-Jan Wagenmakers [c], Ben Blum [d]

[a] *Department of Cognitive Sciences, University of California, 3151 Social Sciences Plaza, Irvine, CA 92697-5100, USA*
[b] *Massachusetts Institute of Technology, Cambridge, MA, USA*
[c] *Northwestern University, Evanston, IL, USA*
[d] *Stanford University, Stanford, CA, USA*

Received 5 May 2002; accepted 31 July 2002

Abstract

Information about the structure of a causal system can come in the form of observational data—random samples of the system's autonomous behavior—or interventional data—samples conditioned on the particular values of one or more variables that have been experimentally manipulated. Here we study people's ability to infer causal structure from both observation and intervention, and to choose informative interventions on the basis of observational data. In three causal inference tasks, participants were to some degree capable of distinguishing between competing causal hypotheses on the basis of purely observational data. Performance improved substantially when participants were allowed to observe the effects of interventions that they performed on the systems. We develop computational models of how people infer causal structure from data and how they plan intervention experiments, based on the representational framework of causal graphical models and the inferential principles of optimal Bayesian decision-making and maximizing expected information gain. These analyses suggest that people can make rational causal inferences, subject to psychologically reasonable representational assumptions and computationally reasonable processing constraints.
© 2003 Cognitive Science Society, Inc. All rights reserved.

Keywords: Causal reasoning; Decision making; Bayesian networks; Bayesian models; Rational inference; Hypothesis testing; Active learning; Observational learning; Interventions; Structure learning; Web experiments; Human experimentation; Computer simulation

* Corresponding author. Tel.: +1-949-824-7642; fax: +1-949-824-2307.
E-mail addresses: msteyver@uci.edu (M. Steyvers), jbt@mit.edu (J.B. Tenenbaum).
[1] Tel.: +1-617-452-2010; fax: +1-617-258-8654.

0364-0213/03/$ – see front matter © 2003 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/S0364-0213(03)00010-7

1. Introduction

The ability to infer causal relationships is crucial for scientific reasoning and, more generally, forms the basis for learning to act intelligently in the world. Knowledge of causal relationships, as opposed to mere statistical associations, gives us a sense of deep understanding of a system and a sense of potential control over the system's states, stemming from the ability to predict the consequences of actions that have not yet been performed (Pearl, 2000). There has been extensive study of how people make inferences about simple causal relationships, focusing on learning the relationship between a single cause and effect (e.g., Anderson, 1990; Buehner, Clifford, & Cheng, 2002; Cheng, 1997; Griffiths & Tenenbaum, 2003; Jenkins & Ward, 1965; Lober & Shanks, 2000; Shanks, 1995; Tenenbaum & Griffiths, 2001; see also Hagmayer & Waldmann, 2000; White, 2000). In the present research, we are interested in how networks involving multiple cause–effect relationships can be inferred. This question has recently received much attention in philosophy and computer science (Glymour & Cooper, 1999; Pearl, 2000, 1988; Spirtes, Glymour, & Scheines, 2000) but has received relatively little attention in psychology. The problem of inferring cause–effect relationships is difficult because causal relations cannot be observed directly and instead have to be inferred from observable cues. In addition, in order to infer the structure of a network of multiple cause–effect relationships, we have to understand how individual cause–effect relationships interact.

We study people's ability to infer the structure of causal networks on the basis of two kinds of statistical data: pure observations and experimental manipulations. The former case involves passive observation of random samples of a system's autonomous behavior, while in the latter case, the learner can actively control some variable in the system and observe the corresponding effects on other variables in the system. The difference between passive and active learning has also been compared to the difference between learning from watching and learning by doing, or the difference between correlational (non-experimental) and controlled experimental studies in science (Pearl, 2000). We will refer to the data gathered from passive and active learning as observational and interventional data, respectively.

Our studies of human causal learning are motivated by the computational framework of learning in causal graphical models, or Bayesnets (Glymour, 2001; Glymour & Cooper, 1999; Pearl, 2000, 1988; Spirtes et al., 2000). The theory of learning in graphical models explains how and when causal structure may be inferred from statistical data, either passive observations, interventions or a combination of the two. We propose computational models of human causal learning in a rational framework (Anderson, 1990; Oaksford & Chater, 1994), based on Bayesian inference over causal graphical model representations. Our models also adopt certain representational and processing constraints, which are not intrinsic to the rational framework but which lend added psychological or computational plausibility.

Our framework for modeling people's intervention choices is inspired by work in statistical machine learning on active learning of causal networks (Murphy, 2001; Tong & Koller, 2001). This work treats the task of intervention selection as an information maximization problem: the goal is to pick the intervention for which, when we observe its effects, we can expect to gain the maximum possible information about the underlying causal structure. Data selection and hypothesis testing strategies have long been studied in scientific discovery (e.g., Klahr & Dunbar, 1988), concept learning (e.g., Bruner, Goodnow, & Austin, 1956) and the reasoning

about rules (Wason, 1968), including recent work from an information-maximization standpoint (Oaksford & Chater, 1994; Nelson, Tenenbaum, & Movellan, 2001). Yet, to our knowledge, this approach has not been pursued previously in the context of human causal learning.

The plan of the paper is as follows. We first give a short overview of statistical approaches to learning causal structure with graphical models. We then present a series of three experiments with human subjects, along with computational models of each task and model-based analyses of the experimental results. We close with a discussion of how this work relates to other studies of human causal inference and some open questions.

2. Structure learning in causal graphical models

This section is not intended as a general introduction to causal graphical models. Its purpose is to explain how causal structure can be inferred on the basis of probabilistic relationships between variables (c.f., Glymour & Cooper, 1999; Pearl, 2000; Spirtes et al., 2000), and to introduce the family of causal networks used in our empirical work. For complementary perspectives on how the paradigm of graphical models may be brought to bear on questions of human causal inference, see Glymour (2001), Gopnik et al. (in press), Tenenbaum and Griffiths (2001, in press), and Griffiths and Tenenbaum (2003).

Directed graphs provide us with an intuitive way to express causal knowledge (see Fig. 1). Nodes represent continuous or discrete state variables of a system and arrows represent direct causal relations, pointing from causes to effects. The state of each variable is modeled as some function of the states of its parents—its direct causes. The functions relating causes to their effects may be probabilistic or deterministic. These functions are assumed to be local and modular (Reichenbach, 1956; Suppes, 1970), operating independently for each "family" of a state variable and its parents. The states of variables with no parents may be determined exogenously or by some stochastic process with a particular prior probability distribution. Together, these ingredients define a joint probability distribution over all n state variables $X_1, \ldots, X_n$:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{parents}(X_i)), \tag{1}$$

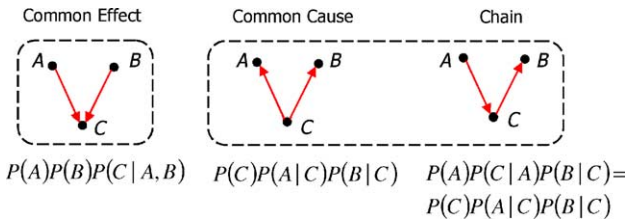

Fig. 1. Three causal network structures, shown with the corresponding factorizations of the joint probability distributions $P(A, B, C)$. Dashed lines group together Markov-equivalent networks, which imply equivalent factorizations of the joint probability distribution and thus are not in general statistically distinguishable based on pure observational data.

where each term $P(X_i|\text{parents}(X_i))$ defines the local causal process for one node X_i as a function of the states of its parents, parents(X_i). Any joint probability distribution can always be expressed in the form of Eq. (1) in many different ways, by choosing any arbitrary ordering of nodes and letting the parent set of node i include all prior nodes $1, \ldots, i-1$ in the ordering. But if the system does have a causal structure, then a causally based factorization will almost surely provide a more compact representation of the joint probability distribution than an arbitrarily chosen factorization.

The factorization of Eq. (1) embodies what is known as the causal Markov condition: the state of any variable is probabilistically independent of its non-descendants given the states of its parents. As a concrete example, consider the three-node chain network shown in Fig. 1, where the only direct causal links are from A to C and C to B. In this network, the variables at the two ends of the chain, A and B, are marginally dependent: knowing nothing else about the system, observations of A will carry information about B, and vice versa. However, conditioned on the observation of the intervening variable C, A has no further influence on B. Thus, the probability distributions of A and B are conditionally independent given C. No other independencies exist for this system; the pairs $\{A, C\}$ and $\{C, B\}$ are always dependent because they are connected by direct causal links. The Markov condition is itself an instance of a very general and useful notion, that the causal structure of a system places constraints on the possible patterns of data that can be observed. Causal inference exploits this principle to reason backward from observed patterns of data to the causal structure(s) most likely to have generated that data.

2.1. Distinguishing networks by observations

The Markov condition directly forms the basis for many algorithms for causal inference (Glymour & Cooper, 1999; Pearl, 2000; Spirtes et al., 2000). These algorithms attempt to estimate the statistical dependencies that hold in a given data set and then to construct the set of causal graphs that, according to the Markov condition, are consistent with just those dependency constraints. Such "constraint-based" algorithms are based on efficient and elegant methods for searching the space of all possible causal structures, but their capacity for causal inference is limited by the limited nature of the constraints they exploit. They look only at whether two variables are statistically dependent, an empirical relationship that could derive from many different underlying causal structures and that may take a fairly large sample size to establish with confidence. They ignore more fine-grained but psychologically salient cues—such as whether one variable is always present when the other is, or never occurs when the other occurs, or always takes on the same value as another variable—which may support much more rapid and confident structural inferences for particular kinds of causal systems.

We consider causal inference more broadly from a Bayesian point of view (Cooper & Herskovits, 1992; Friedman & Koller, 2002; Heckerman, 1999; Heckerman, Meek, & Cooper, 1999; Tenenbaum & Griffiths, 2001, in press). A Bayesian learner considers a set of hypotheses corresponding to different graph structures or different choices of the local probability functions relating causes to effects. Each hypothesis assigns some likelihood to the observed data, and is also assigned a prior probability reflecting the learner's prior knowledge or biases. The posterior probability of each hypothesis—corresponding to the learner's degree of belief that it is the true causal model responsible for generating the observed data—is then proportional to the product

of prior probability and likelihood. While the Markovian learning algorithms of Pearl (2000) and Spirtes et al. (2000) can often be interpreted as asymptotically correct approximations to Bayesian inference, algorithms that directly exploit Bayesian computations may be able to make much stronger inferences from more limited data. Because our experiments—like many situations in the real world—present learners with only a small number of observations, we will base our models on the machinery of Bayesian inference. We save the mathematical details of our algorithms for later in the paper, but first we present an intuitive example of how such reasoning works.

Consider the three example networks shown in Fig. 1: common-effect, common-cause, and chain models. To give an intuitive understanding of the different observations that might be produced by these networks, assume that nodes in the network can be either on or off, that causes are likely to turn their effects on, that in general effects cannot be produced without a cause, but that a node without an explicitly modeled cause (i.e., with no parents in the graph) can turn on spontaneously under some exogenous influence. For example, in the common-effect model of Fig. 1, two causes A and B have the effect C in common. The nodes A and B can turn on spontaneously (and independently of each other), and if either is turned on, it is likely that C will turn on as well. Under this model, likely observations include cases where no nodes turn on, A and C but not B turn on, B and C but not A turn on, or all three nodes turn on. The relative probabilities of these events depend on how likely it is that A or B are activated exogenously. If these base rates are not high, the situation where all nodes turn on is relatively unlikely. In the common-cause network, the arrows are reversed: two effects A and B now have one cause C in common. If C turns on spontaneously, then it probably turns on both A and B as well. Under this model, likely observations include cases where no nodes turn on or where all nodes are turned on. In other words, we expect to see a three-way correlation between all variables—either all on or all off—under the common-cause model, but not under the common-effect model.

Not all differences in causal structure lead to differences in the likelihood of observations. The chain network in Fig. 1—A causes C, and C causes B—has a different causal structure from either the common-effect or the common-cause network. However, like the common-cause network, it tends to produce observations in which either all the nodes are on—when A has been activated exogenously, and thus turns on C which turns on B—or all the nodes are off—when A has not been activated. Thus, a causal chain may not be distinguishable from a common-cause network based purely on passive observations, while either network may in general be distinguished from the common-effect network.

These intuitions can be made precise by manipulating Eq. (1), which expresses the likelihood of different patterns of observation under each causal model. For the common-cause network, Eq. (1) expresses the joint probability distribution as $P(A, B, C) = P(C)\,P(A|C)\,P(B|C)$. For the chain network shown in Fig. 1, Eq. (1) gives $P(A, B, C) = P(A)\,P(C|A)\,P(B|C)$. By applying Bayes' theorem, we can rewrite the chain network likelihood as $P(C)\,P(A|C)\,P(B|C)$, equivalent to the distribution for the common-cause model. Thus, any statistical pattern of observations consistent with one of these network structures can also be generated by the other structure. More formally, these network structures are *Markov equivalent*, meaning that they will in general produce data with the same set of conditional independence and dependence relationships. If two structures are Markov equivalent, then for any way of choosing the local

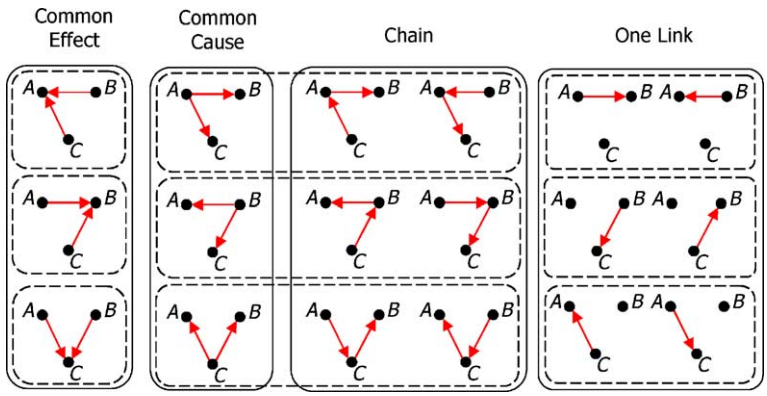

Fig. 2. All three-node causal networks with one or two arrows. Solid lines group together networks of the same topological type. Dashed lines delineate Markov equivalence classes.

probability functions relating causes to effects in one of the networks, there is some way of choosing these functions in the other network that predicts exactly the same distribution of data. Without further knowledge, observational data alone provide no way to distinguish Markov-equivalent structures, such as the common-cause and chain networks.

If we use Eq. (1) to express the joint distribution for the common-effect structure, $P(A, B, C) = P(A)\,P(B)\,P(C|A, B)$, we can see that this structure is not Markov equivalent to either of the other two structures. There is no way to apply the identities of probability theory, such as Bayes' theorem, to express this joint distribution in the same form as the other two. Thus, we can make precise the above intuition that the common-effect structure is in general distinguishable from the other two structures based on purely observational data. We can also extend this analysis to much broader classes of network structures.

Fig. 2 shows all three-node causal networks with either one or two arrows, divided into nine Markov equivalence classes. This set includes all common-cause, common-effect and chain models, along with "one-link" models in which one node is causally disconnected from the rest of the network. Given sufficient observational data, both constraint-based and Bayesian approaches to causal inference can always distinguish between causal models in different Markov classes. Constraint-based methods can never distinguish models in the same Markov class. Bayesian approaches can do so only if they can exploit some particular structure for the local probability functions that breaks the symmetries between cause and effect.

2.2. Distinguishing networks by intervention

When a set of variables comes under experimental control, making it possible to effectively probe a network to test specific causal hypotheses, this is called an intervention. For example, consider the three networks in Fig. 3A. These networks, one common-cause structure and two chain structures, share the same undirected graph but differ from each other in the direction of one or more arrows. These three networks are Markov equivalent, with identical patterns of statistical dependency under observation, but under intervention they become distinguishable.

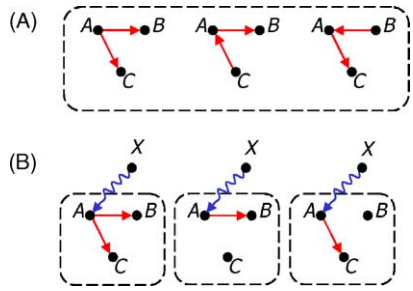

Fig. 3. (A) A Markov equivalence class containing three networks which imply the same statistical dependencies under passive observation. (B) An ideal intervention (X) applied to the middle node, fixing its value to some constant, screens off all other causes of that node. The effect of such an intervention can be modeled as removing any incoming causal links to the target node. Note that the resulting structures on the three original nodes now become distinguishable based on their statistical dependencies; they fall into different Markov classes, as depicted in Fig. 2.

Formally, an intervention on a single variable A can be modeled by inserting an extraneous variable into the causal network as an additional parent of A. In an ideal intervention, the manipulated variable A becomes uniquely determined by the external influence; the intervention screens off the influences any other causes. Graphically, this effect may be represented by "surgery" on the network (Pearl, 2000): all other incoming arrows to the manipulated variable A are deleted. Fig. 3B illustrates how this surgery may break the Markov equivalence between networks that previously had the same patterns of statistical dependency, by rendering variables independent that were previously dependent.

To generate some intuitions about the utility of interventions, let us assume a network that functions as in the previous example, and that intervening on a node turns it on. If we intervene on A, under the hypothesis $C \leftarrow A \rightarrow B$, the likely result is that all nodes would turn on; under the hypothesis $C \rightarrow A \rightarrow B$, only A and B would turn on; and under the hypothesis $C \leftarrow A \leftarrow B$, only A and C would turn on. This intervention is maximally effective for identifying the true causal structure, because it leads to different predictions for all three networks. If instead we intervene on B, the most likely results are the same under the hypotheses $C \leftarrow A \rightarrow B$ or $C \rightarrow A \rightarrow B$—only B would turn on—but quite different under the hypothesis $C \leftarrow A \leftarrow B$, where all nodes would probably turn on. This intervention is potentially diagnostic but not ideal. Therefore, in this example, node A seems to be the most informative target of an intervention. We will later discuss an active learning framework (Murphy, 2001; Tong & Koller, 2001) that formalizes this strategy of choosing interventions based on how much information they can be expected to provide about the underlying structure of the system.

3. Overview of experiments

We will focus on three empirical questions. First, to what extent can people learn causal structure on the basis of purely passive observational data? Because observational data is essentially correlational in nature, this becomes a version of the classic question: to what extent can we infer causation from correlation? Early statisticians famously argued that causation can

only be inferred from randomized experiments, never from correlations (Fisher, 1925), or that there is no need for a notion of causality at all (Pearson, 1911). In parallel, some psychologists have expressed skepticism about the possibility of causal learning based purely on statistical data (Ahn, Kalish, Medin, & Gelman, 1995; Waldmann & Martignon, 1998), while others have argued against the need for specifically causal inference mechanisms as distinct from more general predictive learning procedures (Rogers & McClelland, in press; Shanks, 1995; Wasserman, Kao, Van Hamme, Katagiri, & Young, 1996). Recent work in causal graphical models (Glymour & Cooper, 1999; Pearl, 2000; Spirtes et al., 2000) provides a coherent framework for defining causality and its relationship to probability, such that it is possible to prove conditions under which certain causal hypotheses can be reliably discriminated based on correlational data alone. Constraint-based algorithms turn this logic into an efficient inference procedure (Pearl, 2000; Spirtes et al., 2000), but there is no evidence yet that people do learn causal structure using these purely bottom-up methods. Waldmann and Martignon (1998) write: "In our view, it is unlikely that human learners are good at inducing the causal relations between several interconnected events solely on the basis of covariation information."

In our inference tasks, we assess whether people can infer causal structure from covariational data, but we also provide some additional sources of constraint that reflect the circumstances of natural causal learning. Most importantly, we give people substantive information about the nature of the causal mechanism that can help explain the observed covariation (Ahn et al., 1995). Thus, if two variables are causally connected, our participants can expect not only that their states will be statistically dependent, but more specifically how the state of the effect will depend on the state of the cause. Our cover scenarios also constrain the space of causal network hypotheses people might consider—quite strongly in Experiment 1, but only weakly in Experiments 2 and 3. The Bayesian inference models we develop use these additional sources of knowledge in combination with observed data patterns to choose among competing hypotheses. Our models thus combine the strengths of covariation-based and mechanism-based approaches to causal induction.

The second empirical question is how and to what extent learning from interventions can be combined with, and hopefully improve upon, learning from observations. The advantages of experiment over simple observation have long been recognized within the Western scientific tradition (e.g., Mill, 1874), and we expected that human causal learning would likewise be more effective with interventions than with purely passive observations. It is less clear whether people can integrate interventional and observational data when both kinds of information are available, in order to increase the statistical discriminability of different causal hypotheses. Experiments 2 and 3 explored this question by first giving participants a round of observational trials, followed by a round of intervention trials after which they could refine their hypotheses about causal structure.

The third empirical question is really a cluster of questions concerning the relations between observational and interventional inference, and the principles which guide people's choice of interventions to probe causal structure. In laying out his approach to scientific discovery, Mill (1874) noted that while only experiments can prove causation, pure observation still plays an important role as the natural guide for experimentation. Entering a new domain, scientists often do not know what questions are worth asking, and which experiments worth doing, until they observe a surprising phenomenon along with some other correlated events that might be

potential causes. Like scientists, people might use observations primarily to form hypotheses and interventions primarily to test those hypotheses. In Experiments 2 and 3, we asked participants to choose an intervention after they had indicated their beliefs about causal structure formed on the basis of pure observations alone. This method allows us to test the extent to which intervention choices effectively discriminated between people's competing causal beliefs, and indirectly to assess what those beliefs might be. In Experiment 3 we allowed participants to indicate multiple hypotheses if they were uncertain about the causal relations, and thus we could test explicitly whether interventions were used effectively to reduce uncertainty.

Several criteria guided our choice of experimental tasks and scenarios. First, to test people's ability to construct causal hypotheses for a novel system, many different structures should be plausible, with the possibility of a causal link between any two nodes and no *a priori* bias about the directionality of those links. Second, to test people's capacity for causal inference from probabilistic information, noisy causal connections should be plausible. Third, to minimize the role of low-level attention and memory bottlenecks, correlations should be readily detectable based on only a small number of trials.

These requirements were met by a task of learning about the communication networks of a triad of alien mind readers. On each trial, participants were shown one communication pattern, with each alien thinking of a three-letter non-sense word (consonant–vowel–consonant, e.g., TUS) that appeared in a thought balloon above his head (see Fig. 4). Participants were informed about the mechanism of mind-reading: if alien A reads the mind of alien B, then A thinks of whatever word B is thinking of. If an alien is reading more than one mind, on each trial he randomly picks the contents of one mind to copy. Participants were also informed that the aliens had a limited vocabulary, and that because of lapses of attention, mind-reading occasionally fails. For example, if A and C are both reading B's mind, a sequence of trials might go like this:

A:	TUS	JOF	LIV	RAH	PIF	TUS
B:	TUS	JOF	LIV	DEX	PIF	TUS
C:	TUS	JOF	LIV	DEX	PIF	KUL

Participants were shown the alien communications for several trials and their task was to learn who was the reading the mind of whom. As desired, this copying mechanism could plausibly connect any two nodes in the alien network, in either direction. Although the copying mechanism (like any process of information transmission) could be noisy, the use of categorical variables with many possible states (words) made it possible to detect correlations over just a few trials.

Fig. 4. A sample screen shot from Experiment 1, showing the thoughts of three aliens with mind-reading abilities. Participants have to infer whether the middle alien is reading the minds of the two outer aliens (a common-effect structure) or the two outer aliens are reading the mind of the middle one (a common-cause structure).

4. Experiment 1

The goal of Experiment 1 was to assess whether people can use purely observational data to distinguish between two three-node networks in different Markov equivalence classes. Fig. 4 illustrates what a participant might see on a single trial. The data were generated from either a common-cause structure, with the two outer aliens reading the mind of the middle alien, or a common-effect structure, with the middle alien reading the minds of the two outer aliens.

The precise probabilistic models used to generate the observed communication patterns are as follows. The aliens have a fixed vocabulary of n different words. In the common-cause model ($A \leftarrow C \rightarrow B$), at each trial, a word for the middle node C is first chosen at random from the vocabulary of n words. Then each outer node A and B has an independent chance to copy the content of node C. With probability α, this mind-reading succeeds; the remainder of the time (probability $1 - \alpha$), a word is picked randomly from the vocabulary (so that A or B may still match C with probability $1/n$). In the common-effect model ($A \rightarrow C \leftarrow B$), random words are selected independently for nodes A and B, and node C has independent chances to successfully read the minds of each of its two parents. Each of those mind-reading attempts succeeds with probability α. If C successfully reads both minds, it randomly chooses the contents of one to copy. If C fails to read either mind, which occurs with probability $(1 - \alpha)^2$, a random word is chosen from the vocabulary. In this experiment, $\alpha = 0.8$ and $n = 10$. Mind-reading succeeds most of the time, but occasional mistakes are made, and the aliens have a sufficiently large vocabulary that the probability of getting a match purely by chance is small but not negligible.

Although there are $n^3 = 1{,}000$ different communication patterns that could be observed, it is useful to divide these into four qualitatively distinct sets. These qualitative patterns are shown in Table 1, along with their probability of being observed under the common-cause and common-effect models (with $\alpha = 0.8$ and $n = 10$). A pattern in which all nodes have the same content is most likely under the common-cause model, while a pattern in which two adjacent nodes are equal and one is different is most likely under the common-effect model. Note that the remaining patterns have equal probability under the common-cause and common-effect models and therefore do not discriminate between these models.

Participants were never shown these probabilities, nor given any feedback during the experiment that could allow them to learn these probabilities in a bottom-up fashion. Presenting feedback might also have allowed people to solve this task as a probabilistic categorization task (Gluck & Bower, 1988), merely by associating the presence or absence of different data patterns with the corresponding correct answers. Without such feedback, participants can only

Table 1
The four data patterns (d) and their probabilities under a common-cause (CC) and common-effect model (CE) with $\alpha = 0.8$ and $n = 10$

	d	$P(d\|CC)$	$P(d\|CE)$
(1) $A = B = C$	All same	0.67	0.096
(2) $A = C, B \neq C$ or $B = C, A \neq C$	Two adjacent same	0.3	0.87
(3) $A = B, A \neq C$	Two outer same	0.0036	0.0036
(4) $A \neq B, B \neq C, A \neq C$	All different	0.029	0.029

solve the task by reasoning about the two causal networks and comparing the observed data with the differential predictions made by these two hypotheses.

4.1. Method

4.1.1. Participants

Forty-seven undergraduate students at Indiana University participated in the experiment. An additional 116 subjects participated via the world-wide web. We will refer to these two groups as lab participants and web participants, respectively. We will not go into detail about the advantages and disadvantages of web experiments since these have been discussed well elsewhere (e.g., Reips, 2002; Reips & Bosnjak, 2001).

4.1.2. Procedure

Participants were informed that the alien communications could follow two patterns, one in which the middle alien reads the minds of the outer aliens and one in which the outer aliens read the mind of the middle alien. Instructions made no reference to causal structure or the terms "common-effect" and "common-cause." The values of the parameters α and n were not given to participants, but they were told that the aliens have a limited vocabulary and that mind-reading works most but not all of the time.

Trials were divided into blocks of eight, with a single structure (common-cause or common-effect) generating all trials within a block. A new structure was chosen randomly for each block. On each trial, participants indicated which structure they believed to be the correct generating model, by clicking one of two response buttons. Each response button described one candidate structure in words and illustrated it using a directed graph drawn on top of the aliens (with arrows pointing from the alien(s) that produced thoughts spontaneously to the alien(s) that copied their thoughts). Participants were given a pre-test to make sure that they understood how these directed graphs captured the directionality of mind-reading. The web experiment and the lab experiment consisted of 20 and 40 blocks of trials, respectively, with each structure occurring on half of these blocks. The common-cause and common-effect blocks occurred in pseudo-random order.

4.2. Results

Because there were many more web participants than lab participants, we first present the results of the web participants and then show how similar results were obtained with the lab participants. In general, for all three experiments reported here, we found results for lab participants and web participants to be remarkably similar, with only one exception in Experiment 3. This similarity attests both to the robustness of our results as well as to the potential usefulness of web experiments.

For each participant, we calculated a simple accuracy measure based on the probability of making a correct response, averaged over all trials on all blocks. The accuracy distribution over all web participants is shown in Fig. 5, top left panel. This distribution clearly shows two modes, one at $P(\text{correct}) = 50\%$, which is chance performance, and another near $P(\text{correct}) = 80\%$. This suggests that there are at least two groups of participants, some who

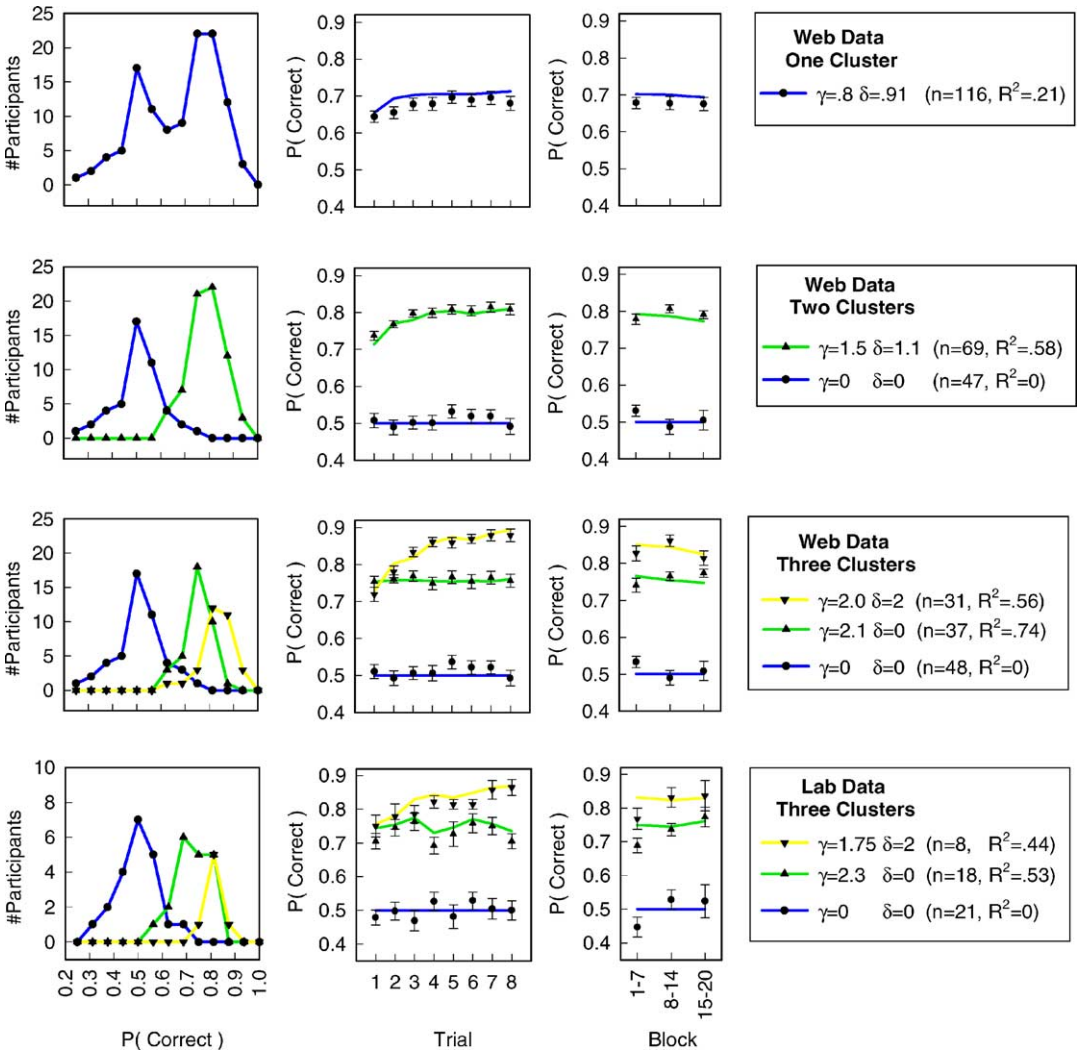

Fig. 5. Results of Experiment 1 analyzed with the model-based clustering method. Rows 1–4 show results for one, two, and three clusters on the web data, and three clusters on the lab data, respectively. Left panels: Accuracy distributions for participants in each cluster. Middle panels: Effects of trial number on mean accuracy in each cluster, averaged over blocks. Right panels: Effects of block on mean accuracy in each cluster, averaged over trials within each block. The legends show, for each cluster, the best-fitting parameter values (γ, δ), the number of participants covered by that cluster (n), and the degree of fit (R^2) between the parameterized model of each cluster and the behavior of participants within that cluster (calculated at the level of individual trials). All panels show error bars for participants' data.

cannot distinguish these two causal models based on observational data and some who can reason relatively accurately about them. The top-middle panel of Fig. 5 shows accuracy as a function of trial number within a block, averaged over blocks. A slight increase in performance over the eight trials indicates that on average, participants are accumulating information

over trials. The top-right panel of Fig. 5 shows accuracy as a function of block, averaged over trials within blocks, with no apparent increase in performance over the course of the experiment.

Because the bimodal accuracy distribution strongly suggests different groups of participants, simply analyzing the average data might not give much insight into the actual causal inference mechanisms used by any one person. Dividing participants into high- and low-accuracy groups, based on some arbitrary threshold, might come closer to revealing how individuals think about this task but would not provide an explanation for these differences. Instead, we examine individual differences in the context of fitting a theoretically based model of causal inference to individual participants' data.

4.3. A Bayesian model for inferring causal structure

The intuition behind our model is as follows. By reasoning about the causal structure of the two candidate networks, people can predict which qualitative data patterns are more or less likely under each hypothesis. People then choose one or the other network in proportion to the weight of this statistical evidence in the data they observe. We can formalize this computation using the log odds form of Bayes' rule:

$$\varphi \equiv \log \frac{P(CC|D)}{P(CE|D)} = \log \frac{P(CC)}{P(CE)} + \log \frac{P(D|CC)}{P(D|CE)}. \tag{2}$$

The log posterior odds, $\varphi = \log P(CC|D)/P(CE|D)$, measures the learner's relative belief in the common-cause (CC) model over the common-effect (CE) model given the data, D. The prior odds, $P(CC)/P(CE)$, expresses the relative degree of belief prior to observing any data. In our experiment, the two networks are equally likely *a priori*, so we will drop the priors from all further calculations. The data D consist of the word patterns $\{d_1, \ldots, d_T\}$ that have been observed from Trial 1 to the current trial T. Assuming that each trial is an independent sample from the true network, we can rewrite Eq. (2) in terms of a sum of single-trial log likelihood ratios:

$$\varphi = \sum_{t=1}^{T} \lambda_t = \sum_{t=1}^{T} \log \frac{P(d_t|CC)}{P(d_t|CE)}. \tag{3}$$

We assume that the likelihoods $P(d_t|CC)$ and $P(d_t|CE)$ are calculated based on the true values of the causal parameters α and n. Table 1 gives the relevant probabilities. For example, suppose that we have observed three samples from a causal model: on two trials, $A = B = C$ ("All same"), and on one trial, $A = C$ but $B \neq C$ ("Two adjacent same"). Then $\varphi = \log(0.67/0.096) + \log(0.67/0.096) + \log(0.3/0.87) = 2.82$; because $\varphi > 0$, the evidence favors the CC model.

To enable this optimal decision model to be compared quantitatively with a broad spectrum of human behavior, and to increase the model's psychological plausibility, we extend it with two free parameters. First, we allow variability in the degree to which people apprehend the diagnosticity of the data. For truly optimal causal inference, a decision of "common-cause" should always follow if and only if $\varphi > 0$. We generalize this decision strategy to one that is

based probabilistically on the odds ratio:

$$P(CC) = \frac{1}{1 + e^{-\gamma\varphi}}. \tag{4}$$

When $\gamma = 1$, Eq. (4) amounts to a probability matching strategy. If the odds ratio favors the common-cause hypothesis by a 2-to-1 margin, then the odds of choosing the common-cause structure over the common-effect structure would also equal 2-to-1. In the limit $\gamma = \infty$, Eq. (4) becomes the optimal strategy of responding *CC* whenever $\varphi > 0$, and *CE* when $\varphi > 0$. In the other extreme, $\gamma = 0$, decisions become random, with no relation to the observed data.

Second, we allow variability in the extent to which people integrate evidence over repeated trials. More recent trials may be weighted more highly, either because people find it difficult to remember earlier observations or because they believe that more recent trials carry more reliable information about the current underlying structure. We thus replace Eq. (3) with a sum of log likelihood ratios weighted by an exponential decay function,

$$\varphi = \sum_{t=1}^{T} \left[\log \frac{P(d_t|CC)}{P(d_t|CE)} \right] e^{-(T-t)/\delta}, \tag{5}$$

where the parameter δ controls the rate of information accumulation across trials. If δ is close to 0, the decision on the current trial depends only on the likelihood ratio of the currently presented data, independent of observations on previous trials within the block. If γ is also high, this decision will be optimal with respect to the current data; we refer to this behavior as "one-trial Bayesian." If $\delta = \infty$, there is no decay. Eq. (5) then simplifies to Eq. (3), and all data patterns are weighted equally, as required for fully optimal Bayesian inference. Intermediate values of δ yield decision behavior intermediate between these two extremes.

4.4. A model-based clustering technique

A common strategy for fitting a parameterized model to individual participant data is to find the best-fitting parameter values for each individual's data. Here we adopt an alternative fitting technique, in which participants are divided into clusters and a single setting of the free parameters describes the behavior of all participants within a given cluster. This procedure allows us to identify subgroups of participants who appear to employ qualitatively distinct approaches to causal inference.

The clustering problem requires solving for two coupled sets of unknown quantities: the assignment of participants to clusters and the best-fitting model parameters (γ, δ) for each cluster. Our fitting algorithm is inspired by *K*-means clustering (see Duda & Hart, 1973). Beginning with random assignments of participants to clusters, we repeat the following two steps until no participant switches clusters:

1. Given the current assignments of participants to clusters, find the parameter settings that best fit the performance of all participants within each cluster.
2. Given the current parameter settings, assign each participant to the cluster whose parameter values fit best.

We assess the degree of fit for a particular parameter setting in a given cluster based on a sum-of-squared-error measure between the model's decisions and each participant's decisions (*CC* or *CE*), summed over all trials and all participants within the cluster. To make predictions for a single individual within a cluster, the model was presented with the exact same sequence of data which that individual saw trial-by-trial during the experiment.

4.5. Model results

Fig. 5 shows the results of the model-based clustering method when one, two, or three clusters were assumed for the web data, and also when three clusters were assumed for the lab data. The plots show accuracy distributions for all participants in each cluster, as well as how mean accuracy depends on trial and block, respectively, separated for each cluster.

4.5.1. Web data, one cluster

When only one cluster is used, the model finds the best-fitting parameters for the whole group of subjects. This parameter fit involves a small amount of decay δ and a scaling parameter γ near one, suggesting that on average, people appear to accumulate some information over trials and to weight likelihood ratios reasonably on this task.

4.5.2. Web data, two clusters

With two clusters, the model locks on to the two modes of the accuracy distribution. It is not equivalent to a hard threshold on accuracy; rather, it divides participants into groups with somewhat overlapping accuracy distributions. Looking at the effect of trial number shows clearly the basis for the model's clustering. One group of participants performs at chance due to a lack of sensitivity to the observed contingencies (i.e., their scaling parameter γ is exactly zero). The other group of participants performs much better, due to a higher scaling parameter and some accumulation of information across trials.

4.5.3. Web data, three clusters

With three clusters, the model splits the good performers of the two-cluster solution into two subgroups. One group (upward pointing triangles) behave as "one-trial Bayesians": they perform relatively well but use only the current data pattern to make their decisions. The other group (downward pointing triangles) approaches optimal Bayesian performance. Note that all three groups are distinguished by qualitatively different parameter settings: the bottom group shows both parameters γ and δ equal to 0, the top group shows both parameters clearly much greater than 1, while the middle ("one-trial Bayesian") group shows a sensitivity γ significantly above 1 but a memory persistence δ equal to 0.

4.5.4. Lab data, three clusters

By applying the clustering technique to a different group of participants tested under lab conditions, we can check the robustness of both the experimental results and the model-based analysis. We analyzed only the first 20 (out of 40) blocks from the lab data, for direct comparability with the web data. The same three qualitatively distinct clusters of inference were found, with very similar parameter values.

4.6. Discussion

Through a model-based clustering analysis of participants' data, we identified three distinct causal inference strategies that people appear to use. The strategies may all be thought of in terms of rational inference, but under qualitatively different representational and processing constraints. The best performers integrate information across trials ($\delta \gg 0$) and reliably make the optimal decision as dictated by the likelihood ratio ($\gamma \gg 0$). The worst performers have both of these parameters at or near zero. An intermediate group—the one-trial Bayesians—are reliably sensitive to the likelihood ratio ($\gamma \gg 0$) but do not integrate information across trials ($\delta = 0$). Thus, we can conclude that most people are to some degree able to distinguish different causal structures on the basis of just a few observed data patterns, and that at least some untutored people are able to perform this task almost optimally. In the next experiment, we will further probe people's causal reasoning mechanisms with an expanded set of causal networks.

The clustering analysis requires that the number of clusters be specified *a priori*, raising the question of how to choose the appropriate number of clusters for the analysis. The bimodality of the overall accuracy distribution strongly suggests that at least two clusters are necessary. By adding a third cluster, we were able to further distinguish between two groups of participants that appeared to perform in qualitatively different ways. Regardless of the random initialization conditions, qualitatively similar results were obtained. However, when four clusters were used, many different clustering solutions were obtained depending on the initialization, and no additional qualitatively different patterns of performance emerged. Thus, a three-cluster decomposition seems to provide the clearest insight into the different ways that people may approach this task.

For the lab participants, there was some indication that performance improved during the course of the experiment (Fig. 5, bottom row). Such a learning effect is interesting because it has occurred in the absence of feedback about the correct answer. This unsupervised learning may be explained in terms of some kind of adaptation to the statistics of the data patterns, or as a re-evaluation of the weight of evidence provided by different data patterns in favor of alternative causal hypotheses. Without any unsupervised learning mechanism, the model we described above does not predict this learning effect. We have developed extensions of the model that do incorporate such learning mechanisms, but for reasons of space, we present here only the simplest version of the model that allows us to pick out individuals' causal inference mechanisms. These learning processes may be a crucial component of real-world causal inference and should be explored more thoroughly in future work.

5. Experiment 2

Participants in Experiment 1 showed reasonable success at inferring causal structure from purely passive observational data. Yet performance for many participants was far from optimal, and the task was relatively simple, with only two possible causal models to consider. In Experiment 2, we compared this passive mode of causal inference based on pure observations with active learning, in which variables come under the learner's experimental control. Participants

reported their beliefs about causal structure twice for every network: once after seeing a round of passive observation trials, and then again after seeing a round of intervention trials, in which one variable that they chose was clamped to a distinctive value and they could observe the effects on other variables in the system. This method enabled us to analyze both the relative efficiencies of inference under passive observation and active intervention conditions, as well as the specific impact of seeing the results of one's own intervention on revising a causal hypothesis.

We also enlarged the set of alternative causal models to include all 18 networks shown in Fig. 2, and we allowed participants to describe their causal hypotheses by a free response method—drawing a network with a mouse—rather than selecting from just two given choices.

Based on observational trials alone, it is impossible to distinguish networks within the same Markov equivalence class (grouped by dashed lines in Fig. 2). But when learners are allowed to intervene, the networks within an equivalence class may become statistically distinguishable. As explained earlier, the degree to which an intervention allows one to distinguish Markov-equivalent networks depends upon the choice of intervention. For all networks in Fig. 2, a single intervention following a round of pure observation trials may be sufficient to identify the true network generating the data, but only if that intervention is well chosen. By allowing participants to choose only a single intervention, we could assess how well they understood the informational value of different choices.

As in Experiment 1, we used the alien mind-reading cover story. The interventions in this experiment are ideal interventions—the manipulation screens off any other causal influences—and were instantiated in the cover story by a special machine (the "mind zapper") that could control a single alien's thoughts and force it to think the word "ZZZ." Fig. 6 illustrates the experimental set-up: following a round of 10 passive observation trials, a participant hypothesizes a certain causal structure (Fig. 6A); then, the participant chooses one alien as the target of the mind zapper and observes the results of 10 more trials while the mind zapper is in place (Fig. 6B). The word "ZZZ" was chosen to be maximally distinct from the normal vocabulary

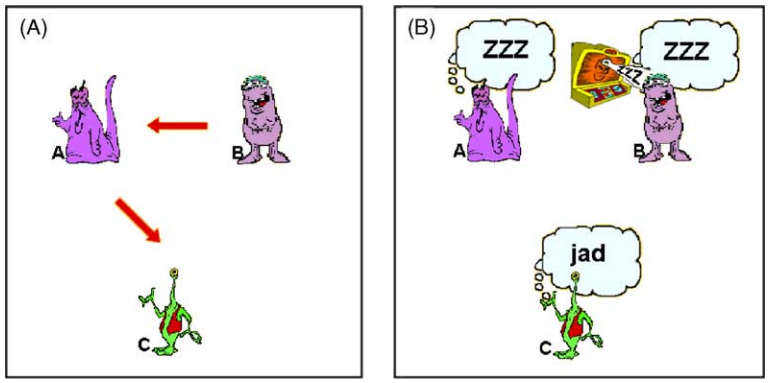

Fig. 6. Illustration of two phases in Experiment 2. (A) In Phase 2 (and then again in Phase 5), participants indicated their hypothesized causal network by clicking arrows onscreen with the mouse. (B) On intervention trials (Phase 4) the "mind zapper" is applied to one alien chosen by participant, fixing the word "ZZZ" in its mind. Observing how that thought propagates to other nodes of the network allows participants to infer the hidden causal structure.

of the aliens, to highlight its role in the observed communication patterns. Once an alien is forced to think about "ZZZ," the same rules for thought transmission apply as in Experiment 1. For instance, in Fig. 6B, the mind zapper is applied to *B* and the result is that both aliens *A* and *B* are thinking about "ZZZ." These data imply a causal connection from *B* to *A*, but do not rule out other possible connections. The network structure in Fig. 6A is still a possibility—it could be that *C* failed to read the mind of *A* by chance. Additional trials might clarify whether a connection exists between *A* and *C*.

5.1. Modeling intervention choice by active learning

We modeled participants' intervention choices within an active learning framework (Murphy, 2001; Tong & Koller, 2001), where learners choose interventions that are expected to provide maximal information about the underlying causal structure. Let g index the possible graphs, D denote all past observational data (on which the intervention decision will be based), a denote the intervention action, and y denote the possible outcomes that could be observed following the intervention. An intervention that maximizes information gained about the true graph is equivalent to one that minimizes the learner's uncertainty H about the true graph following the intervention a,

$$H(a) = -\sum_{g} P_s(g|y,a) \log P_s(g|y,a), \tag{6}$$

where $P_s(g|y,a)$ denotes the learner's distribution of belief over possible graph structures g after taking action a and observing the resulting data y. The subscript s indicates that this is a subjective distribution, reflecting the learner's beliefs without making a commitment to how they are derived. Since we do not know what outcome y will occur prior to choosing action a, we can only minimize the expected value of H over all y:

$$\langle H(a) \rangle = -\sum_{y} P_s(y|a) \sum_{g} P_s(g|y,a) \log P_s(g|y,a). \tag{7}$$

We compute $P_s(y|a)$ as

$$P_s(y|a) = \sum_{g} P_s(y|g,a) P_s(g|a) = \sum_{g} P(y|g,a) P_s(g), \tag{8}$$

where in the second step we have dropped the conditioning on a in $P_s(g)$ because this quantity reflects the learner's beliefs prior to taking action a, and we have dropped the subscript s from $P(y|g,a)$ under the assumption that the learner uses the objective likelihoods (with knowledge of the true values for parameters α and n). From Bayes' rule, we then have $P_s(g|y,a) = P(y|g,a)P_s(g)/P_s(y|a)$.

Under this analysis, intervention choice is determined strictly by the combination of objective likelihoods $P(y|g,a)$, which are a fixed aspect of the task, and $P_s(g)$, the learner's subjective distribution of beliefs over graph structures prior to taking action a. Applying this analysis requires making choices about $P_s(g)$, because we do not observe this quantity directly in our experimental task; we know only the one hypothesis that participants select (out of 18) as their best guess after the passive observation phase of the task.

The most obvious choice for $P_s(g)$ is to set it equal to the Bayesian posterior $P(g|D)$, reflecting the ideal learner's belief distribution given the data D from the passive observation phase. We will refer to this version as the *rational identification model*, because it picks the intervention that is most likely to identify the true causal structure out of all logically possible structures, based on the optimal integration of passive observation and active intervention data. In general, the ideal posterior $P(g|D)$ will be spread across all structures in the same Markov equivalence class as the true structure. Thus, minimizing $\langle H(a)\rangle$ under the assumption $P_s(g) = P(g|D)$ will select an intervention that optimally discriminates members of the same Markov class, as illustrated in Fig. 3.

There are several reasons why a different choice of $P_s(g)$ might be more psychologically plausible. First, it may not be computationally feasible for people—or any learning machine—to compute the full Bayesian posterior over all logically possible hypotheses. Some approximation may be necessary. Second, once participants have chosen a single hypothesis based on their passive observations, their goal in choosing an active intervention may switch from identifying the true structure to testing their chosen hypothesis. We will consider several *rational test models*, in which $P_s(g)$ is determined by assigning most of the probability mass to the participant's chosen hypothesis and the remainder to a small set of simpler alternatives. These strategies may not always lead to the true structure, but they do test particular causal hypotheses in a rational way. They are also much more computationally tractable. Rather than summing over all possible graph structures in Eqs. (7) and (8), rational tests restrict those sums to only a small number of structures that receive non-zero probability under the subjectively defined $P_s(g)$.

The ideal learner always picks the intervention a that minimizes $\langle H(a)\rangle$. We generalize this strategy to a psychologically more plausible choice rule, where action a is chosen with probability

$$P(a) = \frac{\exp(-\beta\langle H(a)\rangle)}{\sum_i \exp(-\beta\langle H(i)\rangle)}, \tag{9}$$

and β is a scaling parameter. The sum in the denominator of Eq. (9) ranges over all possible intervention actions. At one extreme ($\beta = \infty$), choices always minimize $\langle H(a)\rangle$; at the other extreme ($\beta = 0$), choices are random. In all our simulations, we set $\beta = 12$, yielding almost all-or-none behavior except when differences in $\langle H(a)\rangle$ are very small.

5.2. Method

5.2.1. Participants

Twenty-one undergraduates from Indiana University participated. An additional 102 subjects participated via the world-wide web.

5.2.2. Procedure

Instructions were similar to those in Experiment 1, only participants were now introduced to all 18 causal structures in Fig. 2 that could generate the data. Participants could view a schematic of these 18 possible structures for reference at any time during the experiment. Participants were given careful instructions on how to draw candidate network structures, how to interpret the directionality of arrows, and how to use the mind zapper intervention tool. Pretests ensured

participants understood all these elements prior to the experiment. As in Experiment 1, the probability α of copying correctly was set to 0.8, while vocabulary size n was increased to 40 to enhance the discriminability of alternative networks.

The experiment consisted of 8 or 16 blocks of trials for web or lab participants, respectively. On each block, one causal model generated the data. This causal model was randomly selected by first choosing from four types of networks shown in Fig. 2 (common-effect, common-cause, chain or one-link), and then choosing randomly within that type. Each network type occurred equally often over the course of the experiment.

Each block consisted of six phases. In Phase 1, participants were shown 10 self-paced observational trials randomly generated by the causal model. The words were shown in thought balloons and also listed in a separate text window that maintained a record of all trials, to minimize memory demands. In Phase 2, participants chose the single causal network from those shown in Fig. 2 that best explained their observations. They indicated their hypothesis by clicking arrows on screen with the mouse (Fig. 6A). In Phase 3, participants picked a single node as the target of an intervention. In Phase 4, 10 randomly generated trials were shown from the same causal model, with the mind zapper applied to the chosen alien (Fig. 6B). In Phase 5, participants again drew a single causal network representing their best guess following the intervention trials. In Phase 6, feedback was provided to motivate participants. If they had drawn an incorrect network in Phase 5, the correct structure was now given.

5.3. Results

5.3.1. Accuracy of causal inference

We classified a causal network hypothesis as correct only if it matched exactly the generating causal network. Thus, an answer can still be scored as incorrect if it falls into the correct equivalence class, and is therefore indistinguishable from the correct answer even to an ideal learner. This scoring method was chosen in order to assess the accuracy of inferences before intervention (Phase 2) and after intervention (Phase 5) on the same scale.

Before intervention, the mean frequency of correct causal inferences was 18% for both web and lab participants. For comparison, the average score of an ideal learner (assuming knowledge of α and n) would be 50% correct. The distribution of scores over participants is shown in Fig. 7 (filled circles). Performance ranges widely, but overall people perform much better than chance (5.6%), indicated by dashed lines. Even with just a few observations and a large number of potential causal networks, people are still able to make meaningful causal inferences.

After intervention, mean percent correct scores increased to 33% and 34% for the web and lab participants, respectively. See Fig. 7 (open circles) for the full distribution of scores. Again, chance performance is 5.6%, but optimal performance can now be calculated in several ways. An ideal learner who always chooses the most informative intervention and integrates data from all trials would obtain a score of 100%. However, people do not always choose the most informative intervention, nor are they likely to remember the passive observation data (Phase 1) as well as the intervention data (Phase 4). Optimal performance based on participants' actual intervention choices would be 85% (on average), if

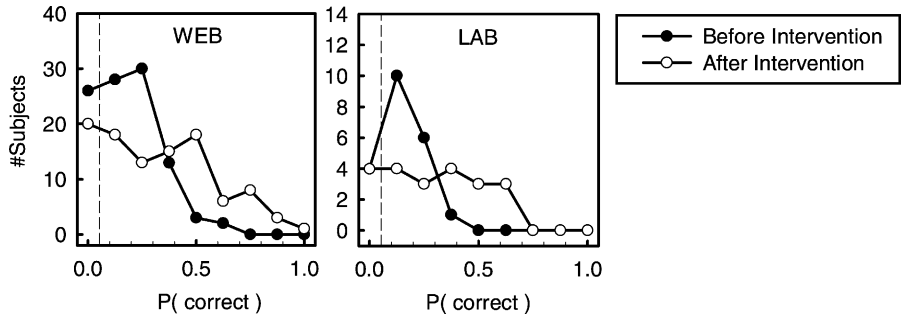

Fig. 7. Distribution of participants' probabilities of choosing the correct causal model in Experiment 2, both before the opportunity to make an intervention (Phase 2) and after seeing the results of their interventions (Phase 5). Dashed lines indicates chance performance (5.6%).

both passive observation and intervention trials are remembered perfectly and integrated, or 53% (on average), if only the intervention data are taken into account in making this final decision.

An analysis of individual participants' choices shows that most people's causal inferences improved following the opportunity to make an intervention relative to their inferences based only on passive observation. Of the web participants, 63 improved in average accuracy following the intervention, 10 became worse, and 29 showed no difference. Of the lab participants, 15 improved, 3 became worse, and 3 showed no difference.

5.3.2. Choice of interventions

Fig. 8 shows the distribution of intervention choices averaged over participants, for both web and lab participants. Choices are conditioned on the type of network selected by the participant after the observational trials, in the previous phase of the experiment. This conditioning allows us to assess how people's intervention choices relate to their hypotheses about causal structure, and the extent to which the interventions serve to test or refine those hypotheses. Choices are classified according to the qualitatively distinct node types in each structure. For instance, given a chain hypothesis $A \rightarrow B \rightarrow C$, there are three qualitatively different choices for intervention: the source node *A*, the mediating node *B*, or the sink node *C*. But given a common-effect hypothesis, $A \rightarrow B \leftarrow C$, there are only two distinct choices: a source node *A* or *C*, or the sink node *B*.

To help interpret the behavioral results, the third row of Fig. 8 shows the predictions of a random choice model, in which nodes are chosen completely at random. Note that because the common-effect network has two source nodes and one sink node, a random choice strategy would generate twice as many interventions on source nodes as on sink nodes given this hypothesis. That random pattern is close to the average choices of participants given a common-effect hypothesis. For all other hypothesis types, people's intervention choices were significantly different from random, with a clear preference for source nodes over sink nodes. Next, we investigate whether this non-random pattern of results can be predicted by rational choice models for active learning of causal structure.

5.3.3. *Active learning—rational identification model*

The fourth row of Fig. 8 shows the predicted results of the rational choice model where the probability distribution $P_s(g) = P(g|D)$, the full Bayesian posterior calculated based

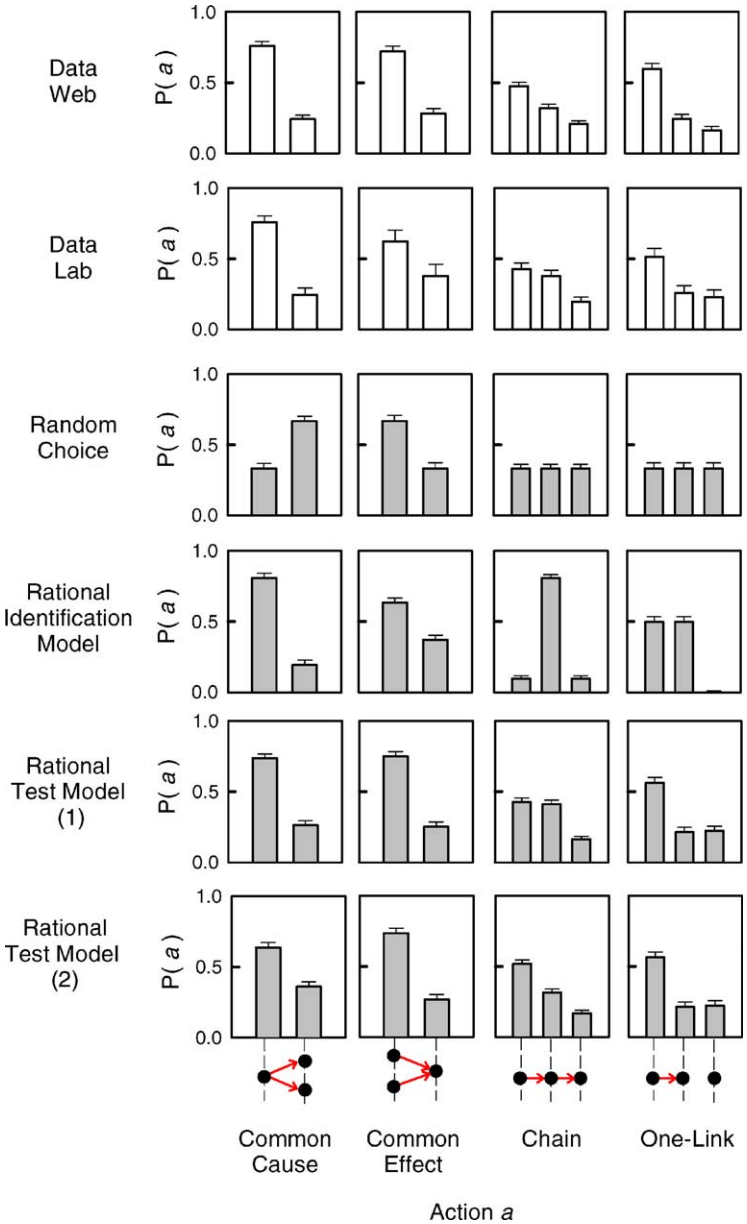

Fig. 8. Distributions of intervention targets chosen in Experiment 2. Intervention choices are conditioned on the type of network selected by a participant in the previous phase, based on passive observations, and classified into topologically distinct node types (source, sink, mediator, or disconnected). The bottom four rows show the predictions of various models explained in the main text. Error bars reflect standard errors.

on all observational trials and assuming perfect knowledge of α and n. This model chooses interventions that, when combined with the passive observation data, are most likely to identify the true causal structure.[1] Predictions are similar to people's choices for common-cause and common-effect networks but are significantly different for the chain and one-link models. There is a crucial difference between these last two cases. People's preference to intervene on the source node given a chain hypothesis is clearly sub-optimal. As illustrated in Fig. 3, choosing the mediating node would be more effective for distinguishing the true structure from alternatives in the same Markov equivalence class. In contrast, people's source preference given a one-link hypothesis is not sub-optimal. It is a kind of bias, because in this case either a source or sink intervention would be equally diagnostic from the standpoint of rational identification. But it is in no way detrimental to successful inference, and in fact represents the most useful intervention people could make here.

In sum, people's source preferences for intervention generally serve the purposes of rational causal inference in this domain, although they are not always predicted by the ideal active learning model, and they can lead to sub-optimal tests in the case of chain-structured hypotheses. We next looked at whether people's intervention choices might be better explained in terms of active learning with a subjective belief distribution $P_s(g)$, defined in terms of the hypotheses that participants report in the previous phase of the experiment based on the passive observation data. These *rational test* models test the current best hypothesis against a small set of alternatives—in contrast to the ideal learner's choice of intervention which attempts to identify the true causal structure from amongst all logically possible structures.

5.3.4. Active learning—rational test model (1)

Perhaps the most basic kind of rational test model assumes that participants aim to test their hypothesized network structure against a "null hypothesis" of complete independence, or no causal structure: $A \rightarrow B \rightarrow C$. While this independence structure never actually occurs in our experiments, it certainly occurs in many real-world situations and may be plausible as a general-purpose null hypothesis for causal learning. To simulate this model, we assigned all of the probability mass in $P_s(g)$ to just two structures: the network h^* that a participant chose after seeing the passive observation data (Phase 2) and the network h_0 with no causal connections. The ratio θ of probability assigned to h^* versus h_0 was set to $\theta = 20$, indicating a much stronger causal belief in the selected network over the hypothesis of complete independence. The model was simulated once for each participant and each block of trials, using the appropriate hypothesis h^* chosen by that participant on that block. Thus, we can compare the distribution of participants' choices (Fig. 8) directly to the distribution of model choices for each network type (aggregated across simulated participants and blocks).

In a further step towards psychological plausibility, we assume that subjects attend only to whether each node shows the distinctive output of the mind zapper, "ZZZ," rather than to the precise pattern of words across all three alien minds. For instance, if a subject intervenes on node *A*, there would be four distinct kinds of observations possible, depending on whether *B*, *C*, neither, or both also show the word "ZZZ."

The predictions of this active learning model come much closer to the distributions of people's choices (Fig. 8, fifth row). The model predicts a source preference for testing common-effect, common-cause, and one-link hypotheses. It also predicts a source preference for testing

chain hypotheses, although the numerical effect is not large. This model's source preference can be explained intuitively. If the null hypothesis of complete independence h_0 is true, any intervention should have the same effect: the target alien will think "ZZZ" and most likely no other alien will. The more aliens who think "ZZZ" in a data pattern, the lower the likelihood of seeing that pattern under h_0. If the hypothesized causal structure h^* is true, intervening on a source node would (for all structures in our task) be most likely to force as many aliens as possible to think "ZZZ"—thereby producing a data pattern which is maximally unexpected under the alternative hypothesis and most diagnostic of the causal structure h^*.

5.3.5. Active learning—rational test model (2)

The final model we explore provides a compromise between considering all logically possible hypotheses and considering only a single null hypothesis of complete independence. Here we assume that learners test their currently favored network hypothesis h^* against all of its sub-networks—those networks with strictly less causal structure. For example, if the learner's current hypothesis corresponds to the chain $A \rightarrow B \rightarrow C$, that hypothesis would be contrasted with $A \rightarrow B \quad C$, $A \quad B \rightarrow C$, as well as the hypothesis of complete independence, $A \quad B \quad C$. Intuitively, this strategy may be seen as attempt to test simultaneously the efficacy of all hypothesized causal links, without worrying about whether additional causal structure might exist. The same parameter setting $\theta = 20$ now controls the relative probability assigned in $P_s(g)$ to the current hypothesis h^* versus the union of all alternative (sub-network) hypotheses, each weighted equally. As shown in the bottom row of Fig. 8, this model comes closest to explaining the preference gradient observed for the chain model, with a substantial preference for the source node over the mediating node, as well as a preference for the mediating node over the sink node.

5.3.6. Model fits

The random, rational identification, rational test (1), and rational test (2) models explain 3%, 11%, 93%, and 93%, respectively, of the variance in the web data, calculated at the level of individual trials. Similarly, these models capture 14%, 11%, 96%, and 85%, respectively, of the variance for the lab data. We do not have sufficient data to quantitatively distinguish the two rational test models.

5.4. Discussion

Our results directly support two conclusions about the role of interventions in causal inference. First, people are more successful at inferring the structure of causal networks after they have had the opportunity to intervene actively on the system and observe the effects, relative to when they have just been exposed passively to pure observations. Second, people's intervention choices can be understood within a rational information-theoretic framework for active learning, in which targets are chosen in proportion to how much information about network structure they can be expected to yield.

The extent to which people's causal inferences approach optimality is a more complex issue. The ideal causal learner, who makes intervention choices by considering all logically possible causal structures and final decisions by integrating all observational and interventional

data according to Bayes' rule, would obtain substantially higher accuracy rates and qualitatively different intervention choices than participants in our study. Yet there is evidence that people do integrate knowledge gained across both observation and intervention trials. Any learner who integrates information from both rounds of data within a block should improve in accuracy—for the ideal learner, from 50% based on passive observation trials alone to 85% based on a combination of observation and intervention trials—while even an ideal learner who did not integrate across both rounds—who approached the intervention trials as if they were a new block of trials—would show no such improvement based on the interventions participants chose. Average performance on our task improved by almost 80% between pre- and post-intervention responses, consistent with substantial integration of knowledge gained from both passive observations and active interventions.

Modeling people's intervention choices clarifies the ways in which these two sources of knowledge might be integrated. Different models of intervention choice differ in the assumptions they make about people's inferential goals and representational constraints. In the rational identification model, the learner's beliefs about the underlying graph structure $P_s(g)$ reflect the full Bayesian posterior, and consequently interventions are optimal with respect to discriminating among members of the Markov equivalence class consistent with the passive observation data. Under this model, data from passive observation and active interventions are integrated on equal terms. However, people's choices appear more consistent with one of the rational test models, in which $P_s(g)$ concentrates on the single favorite hypothesis indicated by a participant, with some small weight allocated to alternative hypotheses with strictly less causal structure. Under these models, passive observation and active intervention play different roles in learning: observation suggests hypotheses, which are then tested through intervention.

The different goals of rational test and rational identification models—testing whether all links in a hypothesized causal structure actually exist, versus attempting to pick out the full causal structure from the sea of all logically possible alternatives—are not ultimately incompatible. Adopting the "test" strategy as a tractable heuristic for causal identification is reminiscent of an approach adopted by many scientists. First, based on observations of a system, make the best possible guess at a strong hypothesis about how that system works. Then, through deliberately chosen interventions, test that strong theory against "less interesting" alternatives in which one or more causal components of the favored hypothesis do not exist, to see if each hypothesized causal component is in fact necessary to explain the system's behavior. Unlike most scientists, participants in this experiment only had the opportunity to try a single intervention. In future work, our active learning paradigm could be extended to study the more general problem of how people plan a "program of research"—a series of interventions, in which the choice of which intervention to perform next depends on the outcomes of previous interventions.

6. Experiment 3

In the previous experiment, participants could indicate their causal beliefs by selecting only a single hypothesis from the set of all possible causal networks. Experiment 3 used the same procedures but allowed participants to select multiple hypotheses if they were uncertain.

This change allows us to test directly people's sensitivity to Markov equivalence classes, by measuring the extent to which people select all and only those networks within an equivalence class when the data lend equal support to the whole class. It also gives us more information about people's subjective distribution over graph structures, $P_s(g)$, allowing us to investigate the extent to which people choose interventions that optimally discriminate among multiple hypotheses they have in mind.

6.1. Method

6.1.1. Participants

Twenty-nine undergraduate students from Indiana University participated. An additional 74 subjects participated via the world-wide web.

6.1.2. Procedure

The procedure used in this experiment was identical to Experiment 2, except for the following differences. Both web and lab participants received 16 blocks of trials. In Phases 2 and 5 of each block, instead of being asked to draw a single causal network, participants were shown a grid of all 18 possible causal networks and instructed to check one or more networks that they believed were consistent with the data seen up to that point. The final feedback phase of each block was omitted.

6.2. Results

6.2.1. Accuracy of causal inference

We computed accuracy scores just as in Experiment 2, with the exception that when m networks were chosen and the correct generating network was among them, we controlled for guessing by counting the score as only a fraction $1/m$ of a correct answer. Average percent correct prior to intervention was 21% and 18% for web and lab participants, respectively, rising to 36% and 30%, respectively, after the intervention trials. Fig. 9 shows the full distributions of scores. As in Experiment 2, the majority of participants improved after being given the opportunity to intervene actively on the system.

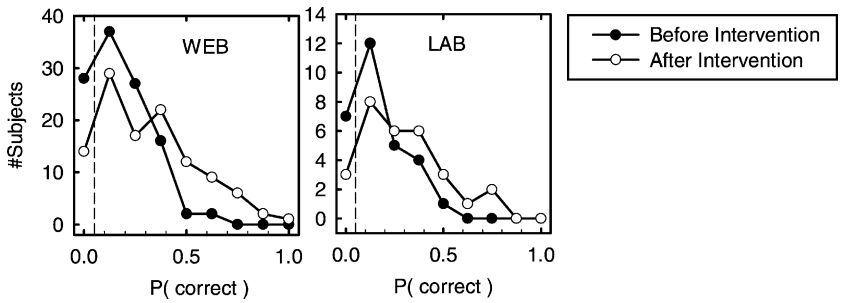

Fig. 9. Distribution of participants' probabilities of choosing the correct causal model in Experiment 3, both before the opportunity to make an intervention (Phase 2) and after seeing the results of their interventions (Phase 5). Dashed lines indicates chance performance (5.6%).

6.2.2. Knowledge of equivalence classes

The matrices in Fig. 10 show $P(j|i)$—the conditional probability of selecting a network from Markov equivalence class j given that a network was also selected from Markov class i—for all blocks on which a participant chose more than one hypothesis. Diagonal entries show the probabilities of pairing a network with another from the same Markov class; off-diagonal entries correspond to pairings across classes. Both web and lab participants appear to understand the one-link Markov equivalence classes: networks within those classes are often grouped together and rarely grouped with networks in other classes. The singleton Markov classes involving common-effect structures are also understood to some extent, although this cannot be seen by comparing diagonal with off-diagonal entries in Fig. 10, because the diagonal entries are zero by definition for these singleton classes. Crucially, the probability of selecting multiple networks when one network had a common-effect structure was only 20%, which is almost half the analogous probability (36%) for the other network types taken together. This significant difference ($p < .01$, two-tailed t-test) suggests some understanding of the fact that each common-effect network belongs to its own equivalence class. Participants have greater difficulty in distinguishing between the Markov classes involving both common-cause and chain networks (columns 4–6). When one network was selected within one of these classes, the other choice was just as likely to fall into any of these three classes.

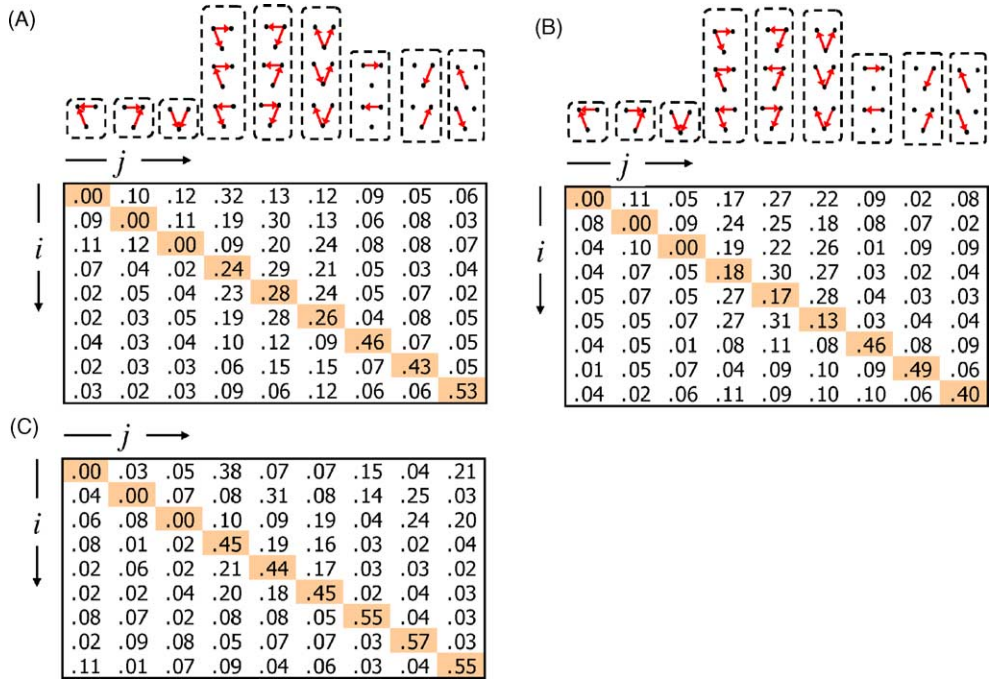

Fig. 10. Conditional probability of participants in Experiment 3 choosing a network from Markov equivalence class j, given that they also chose a network from class i (and that they indicated more than one hypothesis). (A) Web participants. (B) Lab participants. (C) Optimal Bayesian inference, as described in main text.

6.2.3. Analyzing intervention choice

Fig. 11 compares the intervention choices of participants with the predictions of the two rational test models. The data are split into two conditions *post hoc*, based on whether single ($m = 1$) or multiple ($m > 1$) networks were selected after the observational trials. 60% of all subjects chose multiple networks on at least one block. For those subjects, 53% of all blocks resulted in a choice of multiple networks.

Results in the single-choice condition can be compared directly to Experiment 2 and show essentially the same pattern. In the multiple-choice condition, we divided the counts over all the different types of networks that were selected by participants. For example, if on a particular block a participant selected both a common-cause hypothesis and a chain hypothesis, half of the count for that participant's subsequent intervention choice would go to the appropriate

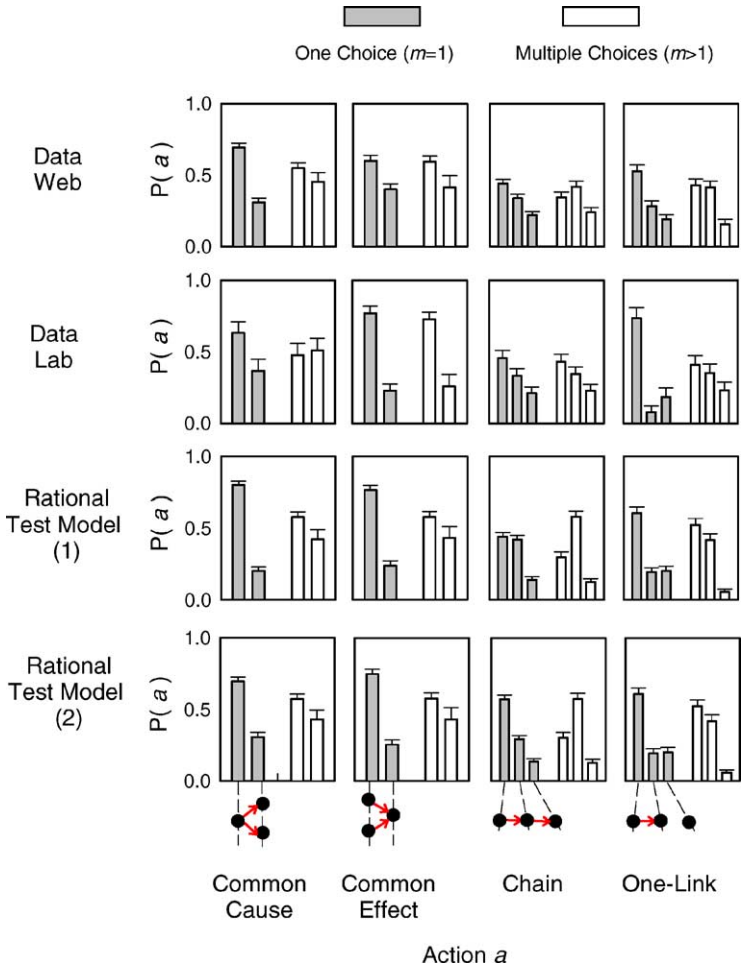

Fig. 11. Distributions of intervention targets chosen by participants in Experiment 3, along with the two rational test models. Results are analyzed separately for trials on which participants chose only a single network hypothesis and trials on which they indicated multiple hypotheses.

type of node (source, mediator, sink) in the common-cause distribution, and the other half to the appropriate type of node (source, sink, unconnected) in the chain distribution. For the rational test models, instead of a single peak in the probability distribution at a single chosen network, the subjective distribution $P_s(g)$ was constructed with m equal peaks in the distribution corresponding to the m chosen networks. The parameter $\theta = 20$ again controlled the relative probability assigned to these m favored hypothesis versus the hypothesis of complete independence (Model 1) or the union of all sub-networks of all m selected networks (Model 2). These two models explained 88% and 91% of the variance, respectively, for web participants, and 82% and 92% of the variance for lab participants, respectively, calculated at the level of individual trials.

6.3. Discussion

Two principal findings emerged from the multiple-choice paradigm used in this experiment. First, causal inference from observational data shows more sensitivity to Markov equivalence classes based on one-link and common-effect structures than to those based on common-cause and chain structures. An analysis of the statistical structure of our task suggests one possible explanation for this difference. For all nine networks in the three Markov classes involving common-cause and chain structures (columns 4–6 in Fig. 10), a single data pattern—where all three aliens think of the same word—is far more common than any alternative observation. Thus, although these three Markov classes are in principle distinguishable purely from passive observations, they are highly confusable given the particular causal mechanisms and the small number of trials used in our task. In contrast, each of the other six Markov classes is clearly and uniquely distinguishable based on its most common one or two data patterns.[2] This argument can be formalized by comparing people's choice probabilities with those of an optimal Bayesian causal learner constrained to choose the same number of hypotheses that participants did (Fig. 10C).[3] Like people, this ideal learner often selects multiple networks across Markov class boundaries when hypothesizing common-cause and chain structures, but rarely when hypothesizing one-link structures.

Our second finding was that when people indicate multiple hypotheses reflecting some uncertainty about causal structure, they often choose subsequent interventions in an attempt to maximally reduce that uncertainty. The close correspondence between human and model choices in this experiment, using the same parameter values as in Experiment 2, further supports the notion that people's intervention choices may be explained as rational tests given their own subjective beliefs about causal structure formed on the basis of passive observation.

Qualitatively, the rational test models appear to fit the data from web participants better than the data from lab participants (except for the $m = 1$ case with common-effect structures, where choices of both people and models are essentially random). Both web participants and the rational test models showed a source preference in every case except one: when attempting to discriminate among multiple hypotheses containing at least one chain structure, both people and the models preferred to intervene on mediating nodes. This exception suggests that the source preference may not be fundamental, but rather may be a consequence of seeking out maximally informative hypothesis tests. We are currently testing this proposal using more complex causal networks in which the source preference and the predictions of rational test models may diverge.

Lab participants did not show this reversal of the source preference, perhaps because they employed less sophisticated strategies. Future work might profitably explore variations in individual strategy on this task, as we did for the simpler two-alternative forced choice task of Experiment 1.

7. General discussion

We have argued that the paradigm of rational statistical inference over graphical models provides insight into multiple aspects of human causal learning: inference from purely observational data, active learning through interventions, as well as the link between these two settings—how passive observations are used to guide interventions toward maximally informative targets. Previous research has investigated Bayesian models of causal inference in more constrained settings (Anderson, 1990; Anderson & Sheu, 1995; Tenenbaum & Griffiths, 2001), and other researchers have recently begun to study empirically people's ability to learn causal networks from observational and interventional data (Lagnado & Sloman, 2002; Schulz, 2001; Sobel, 2003), but we believe our work is the first to draw together all these theoretical and experimental threads.

7.1. Autonomous causal inference in adults

Lagnado and Sloman (2002) and Sobel (2003) both studied observational and interventional learning, focusing on the differences between these two modes. Lagnado and Sloman (2002) used networks of three variables connected via a chain structure and realistic cover scenarios. Participants completed 50 trials of either passive observation or active intervention, in which they could freely choose a new intervention on each trial. As in our studies, Lagnado and Sloman found an advantage for intervention over passive observation, but their participants' overall accuracy scores were much lower. 30% of participants in their intervention condition chose the correct structure out of five alternatives—not significantly different from chance. In their observation condition, participants typically (over 60% of the time) chose an incorrect common-effect structure; fewer than 20% chose correctly. This poor performance was likely influenced by prior knowledge evoked by the realistic cover scenarios, which suggested that there might be two potential causes for a common effect.

In our experiments, the cover scenario was specifically chosen to avoid instilling any expectations about the underlying causal structure. This design left our participants free to engage data-driven learning mechanisms, which led to mean accuracy scores significantly higher than chance in both our observation and intervention conditions—as much as six times higher in the best intervention conditions. It also allowed us to model learning processes in detail without the additional complexities of representing prior knowledge. More generally, human causal inference often makes essential use of richly structured prior knowledge, as Tenenbaum and Griffiths (in press) have shown for several cases of inference from passive observations. Exploring the role of prior knowledge in guiding active learning from interventions is an important area for future work.

Sobel (2003) also adopted a cover scenario that did not favor any particular causal structure, and he also obtained accuracy scores substantially better than chance. Participants in his intervention condition (Experiment 1) obtained an average of 66% correct when learning probabilistic structures. This score was more than three times the chance level (in a five-alternative-forced-choice task), and much higher than the average score of 35% for participants in his observation condition. Sobel's studies differed from ours in several respects. He allowed participants to make as many interventions as they desired—an average of approximately 40 per network (Sobel, personal communication)—while we allowed only one distinct intervention, in order to study the mechanisms of intervention choice, and the interaction of observational and interventional learning. Sobel did not address these issues; his focus was on the relative accuracy of learning given identical data obtained by different means: intervention, observation, or observation of another agent's interventions.

Comparing final accuracy scores between our studies, Lagnado and Sloman (2002), and Sobel (2003) is complicated by the fact that our participants were allowed to make only one intervention per network, while participants in the latter two studies made far more distinct interventions—enough to test each possible causal link multiple times. This constraint, combined with our larger number of choice alternatives (18 vs. 5), may have accounted for the lower absolute accuracy scores in our studies relative to Sobel's. An alternative possibility is that our task and scenario somehow present an unnatural domain for causal inference, which would undermine our claim to have discovered something about how people learn successfully about the structure of real-world systems. To rule out this possibility, we conducted two small-scale studies identical to Experiment 2, except that participants could now choose a different target for the mind zapper on each intervention trial. In one follow-up, 19 web participants were given 10 distinct intervention trials on each block, obtaining an average final accuracy of 55% correct. In another follow-up, 22 web participants were given 25 intervention trials on each block, obtaining an average final accuracy of 72% correct. This performance, at least as strong as that observed in other studies which presented participants with more data trials and fewer choice alternatives, suggests that our task domain provides a reasonable model system for studying the mechanisms of successful causal inference.

7.2. Children's understanding of causal structure and the impact of interventions

Recent work of Gopnik and colleagues has shown that even young children can successfully engage in causal learning, making inferences that appear normative from the standpoint of causal graphical models (Gopnik & Sobel, 2000; Gopnik, Sobel, Schulz, & Glymour, 2001; Gopnik et al., in press). Most relevantly, Schulz (2001 summarized in Gopnik et al., in press) has shown that children and adults can infer causal structure by observing the interventions of an experimenter. As in our studies and those of Sobel (2003), Schulz and Gopnik used a causally neutral cover scenario, with the specific goal of removing temporal cues to causal order. An important difference is that their participants did not choose which interventions to make, but merely watched the experimenter as she demonstrated one or more interventions. Thus, their results demonstrate people's appreciation of the inferential relations between intervention and causal structure, but they do not directly address the processes by which people often learn about causal systems in the real world, through autonomous observation and intervention.

7.3. Bayesian models of causal inference

None of the above studies of active causal inference has attempted to explicitly model people's inference processes, as we do here. There has been previous work on Bayesian models of human causal inference (Anderson, 1990; Anderson & Sheu, 1995; Tenenbaum & Griffiths, 2001) in more constrained settings. Those models focused on inferences about individual cause–effect relations, where it is clear *a priori* which variables are potential causes or effects (see also Buehner et al., 2002; Cheng, 1997; Jenkins & Ward, 1965; Lober & Shanks, 2000; Shanks, 1995). The task is typically to estimate the strength of these given causal connections, or the probability that a given connection exists in a particular instance, rather than to infer the structure of a network of interacting causes, as in our experiments, where *a priori* any two variables may be related as cause and effect. Also, previous Bayesian models have focused on passive inferences from pure observation data, rather than inferences from a combination of observation and intervention data, or the processes of selecting informative targets for intervention, as we consider here. Establishing the applicability of rational statistical inference models in these more complex and more general settings is an important step towards bringing theories of human causal learning closer to the cases of greatest real-world interest.

7.4. Related work in other cognitive domains

Human inferences about causal networks have also been studied in the context of categorization tasks (Rehder, 2002; Rehder & Hastie, 2001; Waldmann & Martignon, 1998; Waldmann, Holyoak, & Fratianne, 1995). That work was also inspired by the framework of causal graphical models, but differs from our project in several ways. In the categorization literature, causal relations are typically thought of as connecting static features of concepts, whereas in our studies—as well as those of Lagnado and Sloman (2002), Sobel (2003), and Schulz and Gopnik (2001; Gopnik et al., in press)—causal relations connect events or state variables of a dynamical system. Applications of graphical models in the categorization literature have primarily focused on how knowledge of causal structure is used in categorization, rather than how it is acquired—our focus here.

It is an open question whether the mechanisms for inferring causal structure in dynamical systems that we and others have studied may also be engaged in learning the causal structure of categories. One important difference is that the opportunity to make interventions is probably not available in most natural settings of category learning. Given that interventional data can be much more useful than purely passive observations in inferring the structure of a complex network, we would expect that data-driven learning should not be an easy or typical route to acquiring knowledge of causal category structure. In support of this prediction, a consistent finding in the causal categorization studies of Rehder, Waldmann, and their colleagues is that successful learning of causal structure depends crucially on explicitly providing learners with the appropriate causal schema. Bottom-up inference of causal category structure with no prior knowledge and only passive observations appears to be quite rare.

The goal of maximizing expected information gain, which drives our active learning model, has also been invoked in rational explanations of active learning outside the domain of causality. The first that we know of was Oaksford and Chater's (1994) model of Wason's (1968) selection

task, involving reasoning about conditional rules. More recent areas of applications include concept learning (Nelson et al., 2001), eye movements (Lee & Yu, 2000), and the development of gaze following (Movellan & Watson, 2002). Those applications—with the exception of the last, which has a causal underpinning—differ in an important way from ours. The role of active inference in non-causal domains is typically restricted to selecting data; the goal is to make inference more rapid or efficient by choosing which aspects of a problem to gather information about first. Intervening on a causal system, in contrast, actually alters the data patterns that can be observed. Not only does this have the potential to speed up learning, but it also allows inferences to be made that would be impossible otherwise, such as distinguishing between structures within the same Markov equivalence class.

One intriguing similarity between our active inference models and those of Oaksford and Chater (1994) is that both involve choosing tests that optimally discriminate between a currently favored hypothesis and a small set of "less interesting" alternatives—as opposed to the much larger and computationally intractable space of all logically possible alternatives. In Oaksford and Chater's (1994) model, rules of the form "If p then q" are tested against the null hypothesis that p and q are independent. Similarly, in our rational test models, causal networks are tested against the null hypothesis that all variables are independent, or against their sub-networks—networks with strictly less causal structure. As discussed under Experiment 2, this strategy does not always support optimally efficient identification of the true causal structure, but it appears to perform well in practice. It also mirrors the most common approach to empirical discovery followed in many scientific fields that deal with complex causal networks, such as psychology or biology. It would be of great interest to establish more precisely the relation between the intuitive and scientific versions of this strategy, and to analyze rigorously its efficiency as a method for discovering the structure of complex systems.

7.5. Psychological correlates of rational causal inference

One final open question concerns the manner in which statistical algorithms for causal inference might actually find their expression in human minds. Our models attempt to explain people's behavior in terms of approximations to rational statistical inference, but this account does not require that people actually carry out these computations in their conscious thinking, or even in some unconscious but explicit format. A plausible alternative is that people follow simple heuristic strategies, which effectively compute similar outputs as our rational models without the need for any sophisticated statistical calculations. For instance, in our first experiment, one such heuristic would be to count the number of trials on which the three aliens all think of the same word, and then to choose the common-cause model if and only if the aliens agree on a majority of trials. This strategy comes close to optimal Bayesian performance, and distinguishing between the two would not be easy. A similar heuristic for success in our later experiments might select hypotheses on the basis of how well the observed data match up with the one or two data patterns most commonly generated by each candidate network. As explained in the discussion of Experiment 3, such a strategy applied to our task would generally be successful for inferring common-effect and one-link structures, but given small samples would tend to confuse common-cause and chain structures across Markov class boundaries—just as participants do in our experiments.

For several reasons, we have focused our first efforts on analyses at the level of rational computation rather than these questions of psychological implementation. Most importantly, a rational analysis is necessary to explain how and when people can reliably infer causal structure from different sources of data: pure observation, active interventions, or a combination of the two. Also, the rational framework allows us to study the effects of various processing constraints on the success of causal inference, and to motivate reasonable heuristic models. It remains a priority for future work on causal inference to establish more explicit links between the computational and psychological levels of analysis.

8. Conclusion

Faced with the challenge of inferring the structure of a complex causal network, and given no prior expectations of what causes what, people bring to bear inferential techniques not so different from those common in scientific practice. Given only passive observational data, people attempt to infer a system's underlying structure by comparing what data they see to what they would expect to see most typically under alternative causal hypotheses. Given the opportunity to learn actively from observing the effects of their own interventions on the system, they make more accurate inferences. When constrained in the number of interventions they can make, they choose targets that can be expected to provide the most diagnostic test of the hypotheses they initially formed through passive observation.

The causal basis of this correlation between scientific and intuitive causal inference is not so clear. Does science follow certain norms because it is essentially intuitive causal inference made systematic and rigorous? Or do educated adults adopt more or less scientific ways of thinking because they are explicitly or implicitly taught these norms in our culture? Temporal priority favors the former, because even young children show an appreciation of the proper relations between causal structure and observed data (Gopnik et al., in press), yet influences of culture on causal reasoning are also well documented (Nisbett, Peng, Choi, & Norenzayan, 2001). Perhaps, in addition, there is a hidden common cause. Scientists and everyday learners adopt similar inferential techniques because these techniques provide computationally efficient and effective means for learning about the structure of the world, and the selection pressures favoring rapid and successful causal inference in both science and everyday life are too great to ignore.

Notes

1. For this rational identification model, intervention choices are shown conditional on the type of network that received maximum posterior probability $P(g|D)$ after the passive observation phase. If multiple network types attained the maximum, the counts for $P(a)$ were calculated for each network separately and divided by the number of networks.
2. For each one-link class, almost all trials show the same thought for the two connected aliens but a different thought for the third alien. For each common-effect class, almost all trials show a thought shared by the sink (effect) node and either one or the other source (cause) nodes, but not both.

3. In order to compare this model directly with people's choices, its selection process was yoked to the choices of web participants. Whenever a participant selected *m* networks on a particular trial, the model selected the *m* networks to which it assigned highest probability. Thus, the model was sometimes forced to select more networks that it otherwise would with a maximum probability rule, crossing Markov class boundaries that would otherwise not be crossed.

Acknowledgments

This work was supported in part by a grant from NTT Communication Sciences Laboratory. Much of this research was conducted while the first two authors were in the Department of Psychology, Stanford University. We are grateful to Tom Griffiths, Kevin Murphy, Vin de Silva, David Sobel, David Langado, Steven Sloman, Alison Gopnik, David Danks, and Tamar Kushnir for many useful discussions. The first author would also like to acknowledge the enormous intellectual stimulation of Rich Shiffrin in previous collaborations on Bayesian inference mechanisms in episodic memory that helped to shape the Bayesian modeling framework in this research.

References

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299–352.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., & Sheu, C.-F. (1995). Causal inference as perceptual judgments. *Memory & Cognition, 23*, 510–524.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2002). From covariation to causation: A test of the assumption of causal power. Manuscript under review, *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–347.

Duda, R., & Hart, P. (1973). *Pattern recognition and scene analysis*. New York: Wiley.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Friedman, N., & Koller, D. (2002). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, to appear.

Glymour, C. (2001). *The mind's arrows*. Cambridge, MA: MIT Press.

Glymour, C., & Cooper, G. (Eds.). (1999). *Computation, causation, and discovery*. Cambridge, MA: MIT Press.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*(3), 227–247.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (in press). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*, 1205–1222.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*, 620–629.

Griffiths, T. L., & Tenenbaum, J. B. (2003). *Structure and strength in causal judgments*, submitted for publication.

Hagmayer, Y., & Waldmann, M. R. (2000). Simulating causal models: The way to structural sensitivity. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the cognitive science society* (pp. 214–219). Mahwah, NJ: Lawrence Erlbaum.

Heckerman, D. (1999). A tutorial on learning with Bayesian networks. In M. Jordan (Ed.), *Learning in graphical models*. Cambridge, MA: MIT Press.

Heckerman, D., Meek, C., & Cooper, G. (1999). A Bayesian approach to causal discovery. In C. Glymour & G. Cooper (Eds.), *Computation, causation and discovery*. Cambridge, MA: MIT Press.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*, 1–17.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1–48.

Lagnado, D., & Sloman, S. A. (2002). Learning causal structure. *Proceedings of the twenty-fourth annual conference of the cognitive science society*. Fairfax, VA.

Lee, T. S., & Yu, S. X. (2000). An information-theoretic framework for understanding saccadic eye-movements. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in neural information processing systems* (Vol. 12). Cambridge, MA: MIT Press.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*, 195–212.

Mill, J. S. (1874). A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific evidence (8th ed.). New York: Harper. (First edition published in 1843).

Movellan, J. R., & Watson, J. S. (2002). *The development of gaze following as a Bayesian systems identification problem*. University of California at San Diego, Institute for Neural Computation, Machine Perception Laboratory Technical Report 2002.01, January 29, 2002.

Murphy, K. P. (2001). *Active learning of causal Bayes net structure*. Technical Report. Department of Computer Science, U.C. Berkeley.

Nelson, J. D., Tenenbaum, J. B., & Movellan, J. R. (2001). Active inference in concept learning. *Proceedings of the 23rd conference of the cognitive science society* (pp. 692–697).

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review*, *108*, 291–310.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Palo Alto: Morgan Kaufmann Publishers.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearson, K. (1911). *The grammar of science*. London: A. and C. Black.

Rehder, B. (2002). *Categorization as causal reasoning*, submitted for publication.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*, 323–360.

Reichenbach, H. (1956). *The direction of time*. Berkeley: University of California Press.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256.

Reips, U.-D., & Bosnjak, M. (Eds.). (2001). *Dimensions of Internet science*. Lengerich: Pabst.

Rogers, T. T., & McClelland, J. L. (in press). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Schulz, L. E. (2001). *"Do-calculus": Adults and preschoolers infer causal structure from patterns of outcomes following interventions*. Paper presented at the Second Biennial Meeting of the Cognitive Development Society, October 27, 2001, Virginia Beach, VA.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *48*, 257–279.

Sobel, D. M. (2003). *Watch it, do it, or watch it done: The relation between observation, intervention, and observation of intervention in causal structure learning*. Manuscript submitted for publication, Brown University.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). New York, NY: MIT Press.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland Publishing Company.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. 13). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (in press). Theory-based causal inference. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15). Cambridge, MA: MIT Press.

Tong, S., & Koller, D. (2001, August). Active learning for structure in Bayesian networks. *Seventeenth international joint conference on artificial intelligence* (pp. 863–869). Seattle, WA.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, *124*, 181–206.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual conference of the cognitive science society* (pp. 1102–1107). Mahwah, NJ: Lawrence Erlbaum.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Wasserman, E. A., Kao, S.-F., Van Hamme, L. J., Katagiri, M., & Young, M. E. (1996). Causation and association. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *Psychology of learning and motivation: Causal learning*. San Diego: Academic Press.

White, P. A. (2000). Naïve analysis of food web dynamics: A study of causal judgment about complex physical systems. *Cognitive Science*, *24*, 605–650.